## AMENDMENTS TO THE SPECIFICATION

Please add the following new paragraphs [0037] through [0064] shown below.

[0037] Figures 10A and 10B illustrate a method for clustering a string, the string including a plurality of characters, in accordance with one embodiment of the present invention. The method illustrated in Figures 10A and 10B includes the steps of:

[0038] identifying R unique n-grams $T_{1...R}$ in the string (step 1005);

[0039] for every unique n-gram $T_S$ (step 1010):

[0040] if the frequency of $T_S$ in a set of n-gram statistics is not greater than a first threshold (step 1015:

[0041] associating the string with a cluster associated with $T_S$ (step 1020);

[0042] otherwise:

[0043] for every other n-gram $T_V$ in the string $T_{1...R,\text{ except S}}$ (step 1025):

[0044] if the frequency of n-gram $T_V$ is greater than the first threshold (step 1030):

[0045] if the frequency of n-gram pair $T_S$-$T_V$ is not greater than a second threshold (step 1050):

[0046] associating the string with a cluster associated with the n-gram pair $T_S$-$T_V$ (step 1055);

[0047] otherwise:

[0048] for every other n-gram $T_X$ in the string $T_{1...R,\text{ except S and V}}$ (step 1060):

[0049] associating the string with a cluster associated with the n-gram triple $T_S$-$T_V$-$T_X$ (step 1065);

[0050] otherwise:

[0051] do nothing (step 1035).

[0052] Figures 11A through 11C illustrate a method for clustering a string, the string including a plurality of characters, in accordance with another embodiment of the present invention. The method illustrated in Figures 11A through 11C includes the steps of:

[0053] identifying R unique n-grams $T_{1...R}$ in the string (step 1105);

[0054] for every unique n-gram $T_S$ (step 1110):

[0055] if the frequency of $T_S$ in a set of n-gram statistics is not greater than a first threshold (step 1115):

[0056] associating the string with a cluster associated with $T_S$ (step 1120);

[0057] otherwise:

[0058] for i = 1 to Y (step 1135):

[0059] for every unique set of i n-grams $T_U$ in the string $T_{1...R,\text{ except }S}$ (step 1140):

[0060] if the frequency of the n-gram set $T_S$-$T_U$ is not greater than a second threshold (step 1145):

[0061] associating the string with a cluster associated with the n-gram set $T_S$-$T_U$ (step 1150);

[0062] if the string has not been associated with a cluster with this value of $T_S$ (step 1125):

[0063] for every unique set of Y+1 n-grams $T_{UY}$ in the string $T_{1...R,\text{ except }S}$ (step 1165):

[0064] associating the string with a cluster associated with the Y+2 n-gram group $T_S$-$T_{UY}$ (step 1170).

Please renumber original paragraph number [0037] to paragraph number [0065], as shown below.

~~[0037]~~ [0065] The text above describes one or more specific embodiments of a broader invention. The invention also is carried out in a variety of alternate embodiments and thus is not limited to those described here. For example, as mentioned above, while the invention has been described here in terms of a DBMS that uses a massively parallel processing (MPP) architecture, other types of database systems, including those that use a symmetric multiprocessing (SMP) architecture, are also useful in carrying out the invention. The foregoing description of the preferred embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not by this detailed description, but rather by the claims appended hereto.